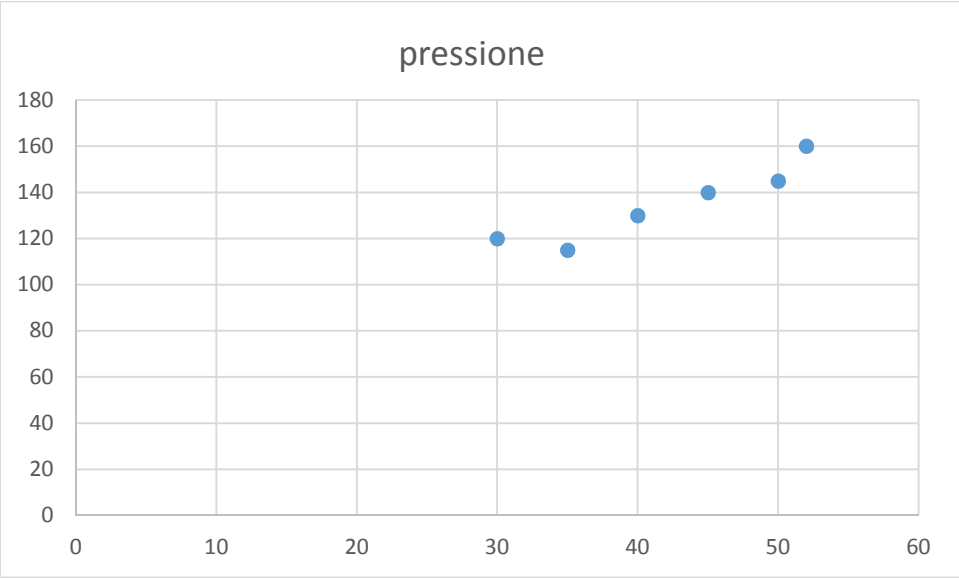
Statistica induttiva

La statistica induttiva (o inferenziale) usa i dati statistici sintetizzati dalla statistica descrittiva per fare previsioni di tipo probabilistico su situazioni future o incerte.
Ad esempio, esaminando un piccolo campione di popolazione, cerca di valutare la frazione di popolazione che possiede una certa caratteristica.
In molti esperimenti interessa l'analisi delle variazioni di due o più variabili per evidenziare eventuali relazioni esistenti tra loro e predire valori non noti sperimentalmente. Considereremo esperimenti con due variabili, la relazione presa in esame è la dipendenza di una variabile rispetto all'altra.
In matematica per esprimere un rapporto di dipendenza, si parla di funzione, in statistica si parla di regressione.
Si indica come indipendente (X) una variabile per cui i livelli possono essere fissati sperimentalmente o possono essere rilevati.
Si indica come dipendente (Y) una variabile la cui variazione è la risposta alle variazioni della variabile indipendente.

Esempio. Nella tabella sono riportate età e pressione arteriose massime (in mmHg) di 6 soggetti maschili.

Età	30	35	40	45	50	52
Pressione	120	115	130	140	145	160

Obiettivo: determinare la dipendenza del valore della pressione (Y) dall'età (X).
I dati si rappresentano nel diagramma di dispersione: esso è costituito da un grafico cartesiano in cui i dati sperimentali sono rappresentati da punti.



Esistono diversi modelli di regressione. Un modello di regressione lineare è quello basato sulla retta dei minimi quadrati.

Osservazione.
Il grafico non basta per capire se effettivamente tra i dati esiste una relazione lineare. Una buona misura della dipendenza è data dal coefficiente di correlazione, indicato con r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Con $\bar{x}, \bar{y}$ medie dei dati x_i, y_i.

Osservazione.

$$\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n}\bar{x}y_i + \sum_{i=1}^{n}\bar{x}\bar{y} =$$

$$= \sum_{i=1}^{n} x_i y_i - n\cdot\bar{x}\bar{y} - n\cdot\bar{x}\bar{y} + n\cdot\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\cdot\bar{x}\bar{y}$$

In definitiva,

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\cdot\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Osservazioni.

- $-1 \le r \le 1$
- Se r è vicino a 1 o a -1, c'è una buona correlazione
- $r = \pm 1 \implies$ correlazione perfetta
- $r < 0 \implies$ correlazione inversa
- $r = 0$ oppure vicino a 0 significa che i dati non sono in relazione lineare

Per calcolare r si può usare la seguente tabella. Consideriamo l'esempio precedente, in cui

$$\bar{x} = 42;\ \bar{y} = 135$$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$x_i y_i$
30	-12	144	120	-15	225	3600
35	-7	49	115	-20	400	4025
40	-2	4	130	-5	25	5200
45	3	9	140	15	225	6300
50	8	64	145	10	100	7250
52	10	100	160	25	625	8320
		$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 370$			$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1400$	$\sum_{i=1}^{n} x_i y_i = 34695$

$$r = \frac{34695 - 6\cdot 42\cdot 135}{\sqrt{370\cdot 1400}} = \frac{675}{\sqrt{518000}} \simeq 0{,}938$$

r è più vicino a 1, quindi c'è una buona correlazione.

I dati possono essere rappresentati dalla retta dei minimi quadrati, di equazione $y = mx + q$, con

$$m = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x}\bar{y}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$q = \bar{y} - m\bar{x}$$

Nell'esempio

$$m = \frac{675}{370} \simeq 1{,}824$$

$$q = 135 - 1{,}824 \cdot 42 = 58{,}39$$

$$y = 1{,}82x + 58{,}39$$

La retta dei minimi quadrati permette di fare previsioni:

- Quale sarebbe la pressione massima di un uomo di 47 anni?

$$y = 1{,}82 \cdot 47 + 58{,}39 = 143{,}93$$

- Quale sarebbe l'età di un uomo con pressione massima pari a 150 mmHg?

$$x = \frac{y - 58{,}39}{1{,}82} = \frac{150 - 58{,}39}{1{,}82} = 50{,}33 \simeq 50 \text{ anni}$$

Probabilità continua

In questo capitolo riprendiamo lo studio della probabilità, in particolare vedremo come studiare con metodi probabilistici esperimenti il cui risultato può assumere valori reali qualsiasi, e non semplicemente un numero finito di valori. Si parlerà quindi di variabili aleatorie continue.
Sia Ω uno spazio campione continuo e sia X una variabile aleatoria su esso definita. Allora X si dice continua.
Il termine "aleatoria" si riferisce al fatto che la variabile può assumere valori diversi in dipendenza dall'esito di un esperimento.

Esempi.
Il risultato del lancio di una moneta o di un dado è una variabile aleatoria discreta.
Il valore di un numero scelto a caso in un intervallo reale oppure la durata nel tempo di un prodotto sono variabili aleatorie continue.

Osservazione.
Se X è una variabile aleatoria discreta, l'insieme dei possibili valori di X è finito o numerabile.
Se X è una variabile aleatoria continua, l'insieme dei possibili valori di X è un intervallo o l'unione di intervalli.

Esempio.
Lancio di una freccia su un bersaglio circolare di raggio $R = 7\ \text{cm}$. Sia P un punto del bersaglio, O il centro del bersaglio, $X = \overline{OP}$ la distanza del punto P dal centro del bersaglio.
X può assumere valori nell'intervallo $[0, 7]$. I valori dipendono dallo strumento usato per misurare:

- Se il righello misura solo i centimetri, allora i valori che può assumere X sono $\{0, 1, 2, \dots, 7\}$, quindi X è una variabile aleatoria discreta. Sia $P_i = P(X = i)$ la probabilità che X assuma il valore i, allora $P(X = i) =$ area del rettangolo.

$\sum aree\ rettangolo = P(X = 0) + P(X = 1) + \cdots + P(X = 7) = P(X \in \{0, 1, 2, \dots, 7\}) = 1$

Si può rappresentare in un istogramma, in cui i rettangoli hanno base pari a 1

- Se il righello misura i mezzi centimetri, $X \in \{0, 0.5, 1, 1.5, 2, \dots, 6.5, 7\}$, quindi X è una variabile aleatoria discreta. L'istogramma è analogo, ma con basi più piccole, pari a 0,5.
- Aumentando la precisione dello strumento di misura, al limite X diventa una variabile continua, $X \in [0, 7]$. L'istogramma diventa il grafico di una funzione continua, f, detta funzione di densità di probabilità.

Definizione.
X è una variabile aleatoria continua se esiste $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ integrabile tale che

1. $f(x) \geq 0, \forall x \in \mathbb{R}$
2. $\int_{-\infty}^{+\infty} f(x)\, dx = 1$
3. $P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(t)\, dt,\ \forall x_1, x_2 \in \mathbb{R}$

La funzione f è detta funzione di densità di probabilità di X.
La relazione

$$P(X \leq x) = P(-\infty \leq X \leq x) = \int_{-\infty}^{x} f(t)\, dt = F(x)$$

Definisce la funzione di distribuzione di probabilità.

Osservazioni.

$$P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1)$$
$$P(X = x_1) = 0$$

Definizioni.
Media:

$$\mu = E[x] = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

Varianza:

$$\sigma^2 = Var[x] = E[(x-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)\, dx$$

Deviazione standard (o scarto quadratico medio):

$$\sigma = \sqrt{\sigma^2}$$

Moda:

Ogni valore x in cui f assume un massimo relativo

Mediana:

$$Med[x] = \tilde{x} \quad \text{tale che} \quad P(X \leq \tilde{x}) = P(X \geq \tilde{x}) = \frac{1}{2}$$

p-esimo percentile:

$$x_{p\%} \text{ è il più piccolo numero } x \text{ tale che } F(x) \leq p\%$$

Esistono diverse funzioni notevoli di densità di probabilità:

- Uniforme
- Binomiale
- Di Poisson
- Normale (o di Gauss)

Noi ci limiteremo ad esaminare la distribuzione normale

Distribuzione normale (o Gaussiana)

La funzione di densità di probabilità è

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}$$

Dove μ, σ rappresentano rispettivamente la media e lo scarto quadratico medio.
Diremo che X è una variabile aleatoria continua con distribuzione normale di media μ e varianza σ^2 e lo indicheremo con

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Studio e grafico di $f(x)$

$$D_f = \mathbb{R}$$
$$f(x) > 0 \quad \forall x \in \mathbb{R}$$

$f(x)$ simmetrica rispetto a $x = \mu$

$$\lim_{x \to \pm\infty} f(x) = \lim_{t \to \pm\infty} e^{-t^2} = 0 \implies y = 0 \text{ asintoto orizzontale}$$

$$f'(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \left[-\frac{1}{2\sigma^2} \cdot 2(x-\mu)\right] \geq 0 \iff -(x-\mu) \geq 0$$
$$\iff x - \mu \leq 0 \iff x \leq \mu$$

La funzione ha un punto di massimo in

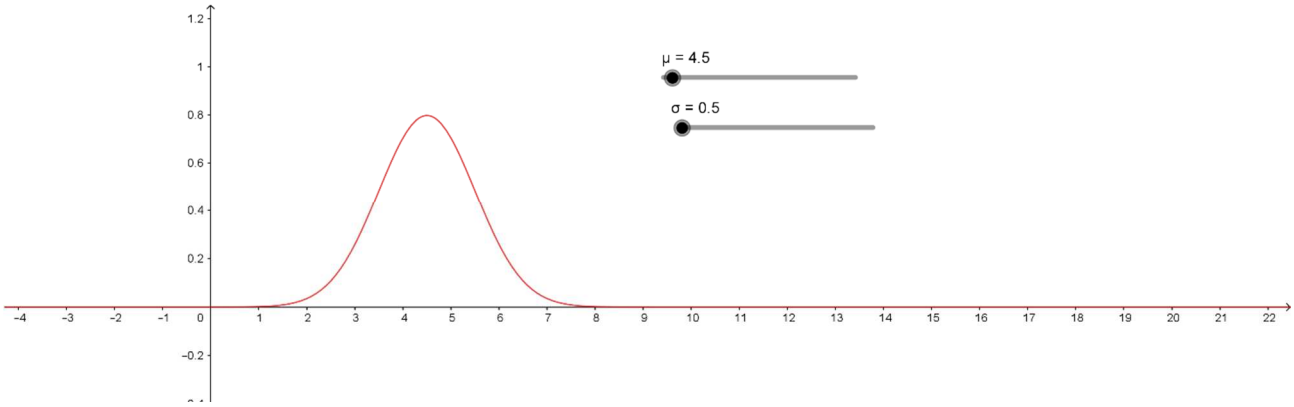
$$\left(\mu\,, \frac{1}{\sqrt{2\pi}\,\sigma}\right)$$

$$f''(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \left[-\frac{1}{\sigma^2} + \frac{(x-\mu)^2}{\sigma^4}\right] \geq 0 \iff \frac{1}{\sigma^2}\left[\frac{(x-\mu)^2}{\sigma^2} - 1\right] \geq 0$$
$$\iff \left[\frac{(x-\mu)^2}{\sigma^2} - 1\right] \geq 0 \iff (x-\mu)^2 \geq \sigma^2 \iff x \leq \mu - \sigma \vee x \geq \mu + \sigma$$

La funzione ha due punti di flesso in

$$\left(\mu - \sigma, \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\,\sigma}\right); \left(\mu + \sigma, \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\,\sigma}\right)$$

Il grafico della distribuzione normale è una curva a campana, detta “gaussiana”



Esempi di caratteristiche che si distribuiscono normalmente:

- Errori di misurazione di una grandezza fisica
- Peso, altezza di una popolazione omogenea
- Dimensione di oggetti prodotti in serie

Osservazione.

$$P(X \leq x) = F(x) = P(-\infty \leq X \leq x) = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

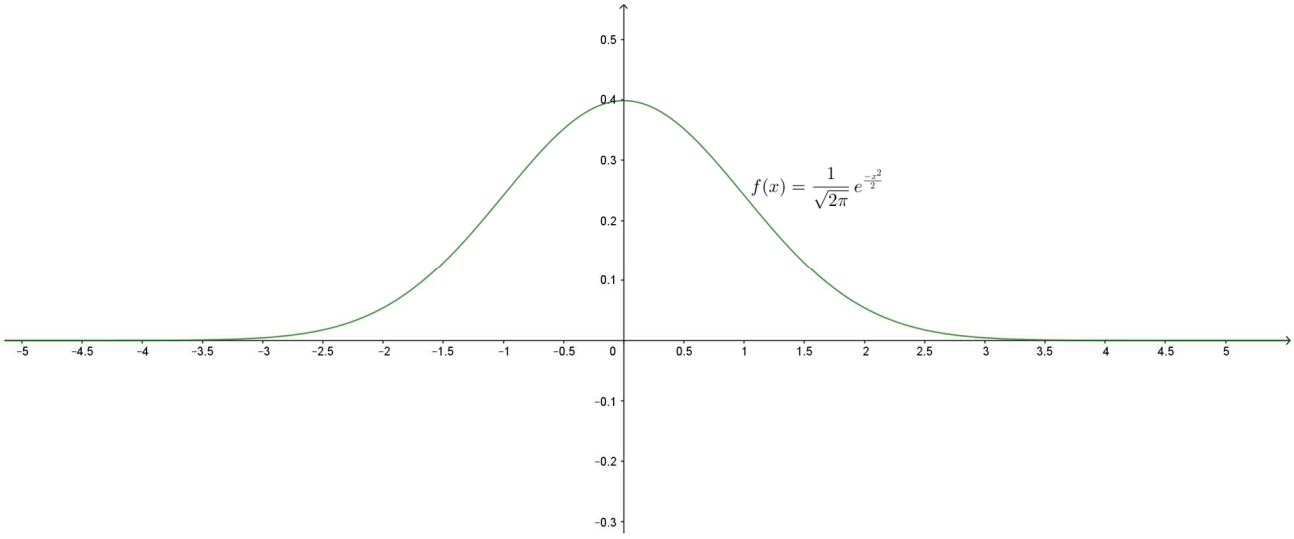
Si calcola utilizzando apposite tabelle in cui è riportato il valore della funzione di distribuzione normale standardizzata.

Al variare di μ, σ esistono infinite distribuzioni normali. Si introduce quindi la funzione di distribuzione normale standard.

Distribuzione normale standard
La funzione di densità di probabilità è

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \forall x \in \mathbb{R}$$

Dove $\mu = 0, \sigma = 1$
$f(x)$ è pari, il suo grafico è quindi simmetrico rispetto all'asse y



In questo caso la funzione di distribuzione $F(x)$ si indica con $\varphi(x)$

$$\varphi(x) = \; P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

Il valore di $\varphi(x)$ si trova nella tabella seguente:

Tabella 1. Funzione di ripartizione della distribuzione normale standard

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Esempi:

$\Phi(0.05) \simeq 0.5199$, $\Phi(1.28) \simeq 0.8997$, $\Phi(3.51) \simeq 1.0000$

$\Phi(-2.41) = 1 - \Phi(2.41) \simeq 1 - 0.9920 = 0.0080$, $\Phi(-4.03) = 1 - \Phi(4.03) \simeq 1 - 1 = 0.0000$

$\Phi(x) = 0.651 \implies x \simeq 0.39$, $\Phi(x) = 0.9921 \implies x \simeq 2.415$,

$\Phi(x) = 0.374 \implies 1 - \Phi(-x) = 0.374 \implies \Phi(-x) = 1 - 0.374 = 0.626 \implies -x \simeq 0.32 \implies x \simeq -0.32$

Regole generali

$$P(X \leq x) = \varphi(x)$$

$$P(X > x) = 1 - \varphi(x)$$

$$P(X \leq -x) = \varphi(-x) = 1 - \varphi(x)$$

$$P(-x \leq X \leq x) = \varphi(x) - \ \varphi(-x) = \varphi(x) - \ \big(1 - \varphi(x)\big) = 2\varphi(x) - 1$$

Dalle tavole si ottiene che:

$$P(X \leq 0) = \frac{1}{2}$$

$$P(-1 \leq X \leq 1) = 2\varphi(1) - 1 \simeq 0{,}6826$$

$$P(-2 \leq X \leq 2) \simeq 0{,}9545$$

$$P(-3 \leq X \leq 3) \simeq 0{,}9973$$

Il $99{,}7\%$ dell'area sta in $[-3; 3]$, per questo motivo le tabelle riportano solo i valori di X in $[-3; 3]$.

Osservazione.
Se X non è standard, cioè $X \notin \mathcal{N}(0, 1)$, bisogna standardizzarla.

Standardizzazione di X

Sia X una variabile aleatoria continua con $X \sim \mathcal{N}(\mu, \sigma^2)$. Allora

$$Z = \frac{X - \mu}{\sigma}$$

È una variabile aleatoria continua con $\mathcal{N}(0, 1)$ che si chiama standardizzazione di X.
Dunque

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \varphi\left(\frac{x - \mu}{\sigma}\right)$$

Per quanto detto in precedenza,

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = \ P(-1 \leq Z \leq 1) \simeq 0{,}6826$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = \ P(-2 \leq Z \leq 2) \simeq 0{,}9545$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = \ P(-3 \leq Z \leq 3) \simeq 0{,}9973$$

Esercizio.

Il livello di glucosio a digiuno in un gruppo di pazienti non diabetici si distribuisce normalmente con $\mu = 98\ mg/_{100\ ml}$ e $\sigma = 8\ mg/_{100\ ml}$. Determinare:

1. $P(X < 82\ mg/_{100\ ml})$
2. $P(90\ mg/_{100\ ml} < X < 106\ mg/_{100\ ml})$
3. $P(X > 116\ mg/_{100\ ml})$

Osserviamo che $X \sim \mathcal{N}(98, 64)$. Dobbiamo standardizzare X.

1. $$P(X < 82) = P\left(Z < \frac{82-98}{8}\right) = P(Z < -2) = \varphi(-2) = 1 - \varphi(2) = 1 - 0{,}9772 = 0{,}0228$$
$$P(X < 82) = 2{,}28\%$$

2. $$P(90 < X < 106\ ) = P\left(\frac{90-98}{8} \leq Z \leq \frac{106-98}{8}\right) = P(-1 \leq Z \leq 1) = 0{,}6826$$
$$P(90 < X < 106\ ) = 68{,}26\%$$

3. $$P(X > 116) = P\left(Z > \frac{116-98}{8}\right) = P(Z > 2{,}25) = 1 - \varphi(2{,}25) = 1 - 0{,}9878 = 0{,}0122$$
$$P(X > 116) = 1{,}22\%$$

Il teorema del limite centrale

Siano $X_1, X_2, \dots, X_n$ variabili aleatorie indipendenti, aventi tutte la stessa distribuzione di probabilità (quindi stessa media μ e stessa varianza σ^2). Allora

$$\lim_{n\to+\infty} P\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le x\right) = P(\mathcal{N}(0,1) \le x)$$

Questo significa che, anche in una popolazione che non segue il modello gaussiano, le medie campionarie, se calcolate su campioni abbastanza grandi, tendono a distribuirsi secondo una legge gaussiana.
Possiamo approssimare la media di $X_1, X_2, \dots, X_n$ con una variabile aleatoria normale, avente media comune μ e varianza $\frac{\sigma^2}{n}$.
Il teorema permette di formulare delle stime attraverso gli intervalli di confidenza.

Sia X una caratteristica di una popolazione P. Supponiamo che X sia normalmente distribuita con media μ e varianza σ^2. Estraiamo un campione casuale di n elementi di P, siano $x_1, x_2, \dots, x_n$ i valori di X per gli elementi del campione.
Si può pensare di avere $X_1, X_2, \dots, X_n$ variabili aleatorie continue con $\mathcal{N}(\mu, \sigma^2)$ e che $x_1, x_2, \dots, x_n$ siano i valori assunti da $X_1, X_2, \dots, X_n$ dopo aver svolto l'esperimento.
Consideriamo la media campionaria

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Per il teorema del limite centrale

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to+\infty]{} \mathcal{N}(0,1)$$

Ossia, per n sufficientemente grande si ha che

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Usiamo queste informazioni per ottenere stime di tipo probabilistico per la media μ.
Sia $p \in (0,1)$ una probabilità; possiamo costruire un intervallo $]\mu_1, \mu_2[$, detto intervallo di confidenza, tale che $P(\mu \in]\mu_1, \mu_2[) = p$

Esempio.

Sia X una variabile aleatoria continua con $X \sim \mathcal{N}(\mu, \sigma^2), \sigma^2 = 9$. Da un campione casuale di 5 elementi della popolazione si ottiene $\bar{x} = 61$. Determinare $]\mu_1, \mu_2[$ tale che

$$P(\mu \in]\mu_1, \mu_2[\,) = 95\%$$

Lo scopo è quello di trovare un intervallo di confidenza al 95% per la media μ.

Sia Z una variabile aleatoria standardizzata, allora

$$P(-z \leq Z \leq z) = 0{,}95$$

Sappiamo inoltre che

$$P(-z \leq Z \leq z) = 2P(Z \leq z) - 1$$

Dunque

$$2P(Z \leq z) - 1 = 0{,}95 \implies P(Z \leq z) = \frac{0{,}95 + 1}{2} = 0{,}975$$

Nella tabella il valore $0{,}975$ corrisponde a $z = 1{,}96$, quindi possiamo scrivere la nostra condizione nel seguente modo:

$$P(-z \leq Z \leq z) = P(-1{,}96 \leq Z \leq 1{,}96) = P\left(-1{,}96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1{,}96\right) =$$

$$= P\left(-1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right) = P\left(-1{,}96 \cdot \frac{\sigma}{\sqrt{n}} - \bar{x} \leq -\mu \leq 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} - \bar{x}\right)$$

Moltiplicando per -1 cambiano tutti i segni della disequazione e anche il verso, quindi la condizione diventerà:

$$P\left(\bar{x} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

Sostituiamo i dati del problema al posto di $\bar{x}, \sigma, n$

$$P\left(61 - 1{,}96 \cdot \frac{3}{\sqrt{5}} \leq \mu \leq 61 + 1{,}96 \cdot \frac{3}{\sqrt{5}}\right)$$

$$P(58{,}37 \leq \mu \leq 63{,}63) = 0{,}95$$

Dunque, al 95%

$$\mu \in]58{,}37; 63{,}63[$$

Osservazione.

Più è grande p, più è ampio l'intervallo $]\mu_1, \mu_2[$. Infatti, se fosse

$$P(\mu \in]\mu_1, \mu_2[\,) = 98\% = 0{,}98$$

Avremmo

$$P(-z \leq Z \leq z) = 0{,}98 \Leftrightarrow z = 2{,}32$$

$$\bar{x} - z \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z \cdot \frac{\sigma}{\sqrt{n}}$$

$$P\left(61 - 2{,}32 \cdot \frac{3}{\sqrt{5}} \leq \mu \leq 61 + 2{,}32 \cdot \frac{3}{\sqrt{5}}\right) = P(57{,}89 \leq \mu \leq 64{,}11)$$

Dunque, al 98%

$$\mu \in]57{,}89; 64{,}11[$$